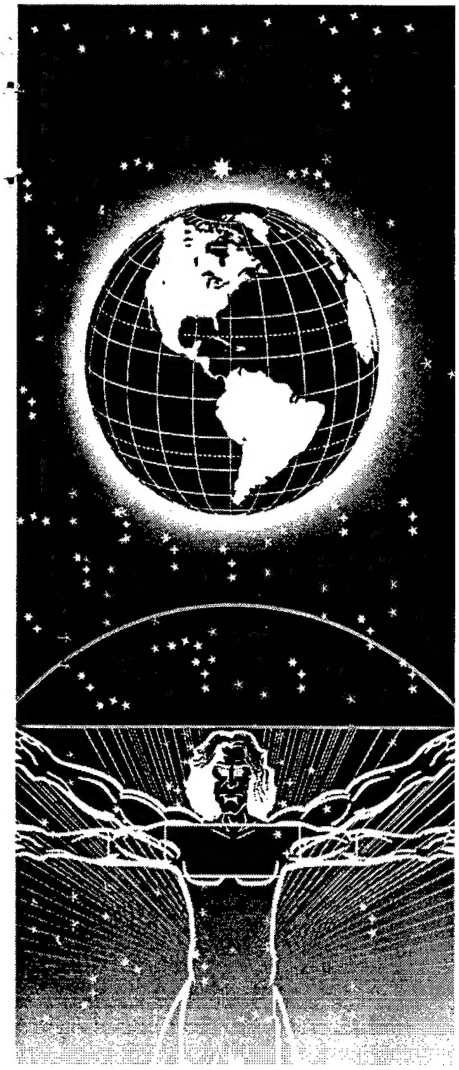
AFRL-HE-AZ-TP-2000-0010



UNITED STATES AIR FORCE RESEARCH LABORATORY

USING THE THORNDIKE MODEL TO ASSESS THE FAIRNESS OF COGNITIVE ABILITY TESTS FOR PERSONNEL SELECTION

Greg A. Chung-Yan
Steven F. Cronshaw

University of Guelph
Department of Psychology
Guelph, ON N1G 2W1 Canada

MAY 2001

AIR FORCE RESEARCH LABORATORY
HUMAN EFFECTIVENESS DIRECTORATE
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

NOTICE

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange. The views expressed in this paper are those of the authors and do not necessarily reflect official views of the US Air Force or the Department of Defense.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Office of Public Affairs has reviewed this document, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

DONALD L. HARVILLE
Contract Monitor

DEE H. ANDREWS
Technical Director

JERALD L. STRAW, Colonel, USAF
Chief, Warfighter Training Research Division

Federal Government agencies and contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218
http://stinet.dtic.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MAY 2001	Thesis	October 1997 to October 1999

4. TITLE AND SUBTITLE

Using the Thorndike Model to Assess the Fairness of Cognitive Ability Tests for Personnel Selection

5a. CONTRACT NUMBER
C – F41624-95-C-5006

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
63227F

6. AUTHOR(S)

Greg A. Chung-Yan
Steven F. Cronshaw

5d. PROJECT NUMBER
2743

5e. TASK NUMBER
A3

5f. WORK UNIT NUMBER
03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Guelph
Department of Psychology
Guelph, ON N1G 2W1
Canada

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Div
6030 South Kent Street
Mesa AZ 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-AZ-TP-2000-0010

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views expressed in this unedited reprint of a thesis are those of the authors and not necessarily those of the US Air Force or the Department of Defense.

14. ABSTRACT

This study evaluates cognitive ability tests (CATs) as predictors of job performance against the Thorndike (1971) model of fairness. Meta-analytic results indicate that CATs substantially misrepresent the relative qualifications between Blacks and Whites in the United States: CATs predict an average job performance difference between groups as three times larger than is actually the case. In practice, then, Blacks are disproportionally burdened by more false-negative selection errors, and this tendency increases markedly under higher CAT cutoffs. Thus, CATs work against proportionate representation of Blacks in the workplace. From an Employment Equity (E.E.) perspective, this is not justifiable because the pool of qualified Black candidates, relative to Whites, is considerably larger than is suggested by CAT scores.

15. SUBJECT TERMS
CAT; Cognitive ability tests; Fairness; Job performance; Personnel selection; Thesis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr Donald L. Harville
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED	65	19b. TELEPHONE NUMBER *(include area code)* (210) 536-3844; DSN 240-3844

CONTENTS

FIGURES

TABLES

PREFACE

I would like to acknowledge the following people for their help and kindly allowing me access to their data: H. John Bernardin of Florida Atlantic University, Ron Boese and the National Center for O*Net Development, Jeffrey M. Conte of San Diego State University, Marie R. Dalldorf and Don McLaughlin of the American institutes for Research, Cathy L. Z. DuBois of Kent State University, Harold W. Goldstein of the City University of New York, Donald L. Harville of the Air Force Research Laboratory, and K. Michele Kacmar of Florida State University.

This effort was conducted in support of thesis requirements and is being reported under USAF Contract F41624-95-C-5006, Work Unit 2743-A3-03, ICATT Programming Support. The Laboratory Contract Monitor is Donald L. Harville.

USING THE THORNDIKE MODEL TO ASSESS THE FAIRNESS OF COGNITIVE ABILITY TESTS FOR PERSONNEL SELECTION

Cognitive ability is well documented as one of the best single predictors of job performance (e.g., Gottfredson, 1986; Hunter, Schmidt, & Rauschenberger, 1984; Ree & Earles, 1991). This has resulted in a trend in employment testing away from specific ability tests for personnel selection and toward more general measures of intelligence or g. Proponents of g in employment testing (e.g. Gottfredson,1986; Ree & Earles, 1991) contend that the g factor is often a better predictor of success in training and performance on the job than an optimally weighted set of specific, job-related scores. Hunter and Hunter (1984) found this to be true across job families with an estimated mean true correlation of .45 for job proficiency and .54 for training success. In addition to the respectable predictive validity of cognitive ability tests (CATs), it is also considered by supporters to be the best way to assess and classify a large number of candidates in terms of probable job success (Landy, Shankster, and Kohler, 1994).

Unfortunately, CATs are also acknowledged as resulting in adverse impact[1] against Blacks in the United States (for simplicity, Blacks and Whites will always refer to Blacks and Whites in the U.S. unless otherwise specified). This is because Black-White differences of approximately one standard deviation (SD) have consistently been cited on

---

[1]Adverse impact refers to a proportionally lower representation of minorities (or other protected group) in comparison to a majority group. It, however, in no way addresses actual job performance. It is merely an indicator of the relative proportion of minority and majority candidates being selected for jobs on the basis of the test scores, irrespective of observed job performance. Thus, evidence of adverse impact cannot be taken as conclusive support, for or against any inherent bias within a test or as indicative of actual discrepancies between races in their average job performance.

measures of cognitive ability, with Blacks scoring lower than Whites (e.g., Jensen, 1980). Efforts to reduce these differences through alternative item formats, while at the same time retaining a comparable level of reliability and criterion-related validity, have not been very successful (Sackett & Wilk, 1994). Thus, when CATs are used for personnel selection, they virtually guarantee adverse impact against Blacks.

Various articles reviewing bias in CATs have concluded that they are fair in the sense that test scores do not mean something different for Blacks and Whites (c.f., Hunter, Schmidt, & Rauschenberger, 1984; Wigdor & Garner, 1982): A particular score on the test predicts with equal accuracy the same level of job performance for both Blacks and Whites. But, an additional consideration identified by Thorndike (1971) when evaluating the fairness of a testing procedure—and the focus of this study—is whether some groups (usually minority groups) are disproportionately subject to higher false-rejection rates—able workers that are incorrectly rejected for the job based on their actual job performance—than other groups (usually the majority group). The Thorndike or Constant-Ratio model considers a test to be fair if the average difference between groups on the predictor is matched by an equivalent difference (in size and direction) on the criterion. A detailed discussion of the Thorndike model and other methods for examining test fairness will be presented later in this thesis.

Although the disproportionate false-rejection rate against Blacks has been cited as a concern (e.g., Campbell, 1996; Hartigan & Wigdor, 1989), little research exists that empirically tests whether Blacks are, in fact, more heavily burdened by false rejection-rates in comparison to Whites. This meta-analytic study examines whether CATs are

considered fair against the Thorndike model.

In order to understand the rationale for this study, the reader must have some familiarity with the concepts and methods used to investigate test bias. The following, however, is a select overview of the test bias literature and only addresses issues that are directly tied to the current study. A more detailed discussion of the various methodologies can be found in Arvey & Faley (1988) and Cole (1973).

Models of Test Bias and Fairness

Differential validity. Differential validity refers to the difference between validity coefficients of two or more groups on a test. It is essentially a comparison of correlation coefficients between groups. Comparisons are often between minority groups (e.g., women or ethnic minorities) and a reference or majority group (e.g., men or Whites). Of interest is the association between the predictor (e.g., test) and the criterion (e.g., job performance) and whether the degree of association is different for minority and majority groups. If the correlation coefficients differ, it is an indication of a biased test because the test is not predicting the criterion with the same accuracy for both groups. As an example, a common expectation for a biased test is that although it may have a useful degree of predictive validity for the majority group, it has less or no useful degree of validity for minority groups (Hartigan & Wigdor, 1989). In other words, the test predicts job performance less well for minority groups than for the majority group. Figure 1 demonstrates this pattern of test results.

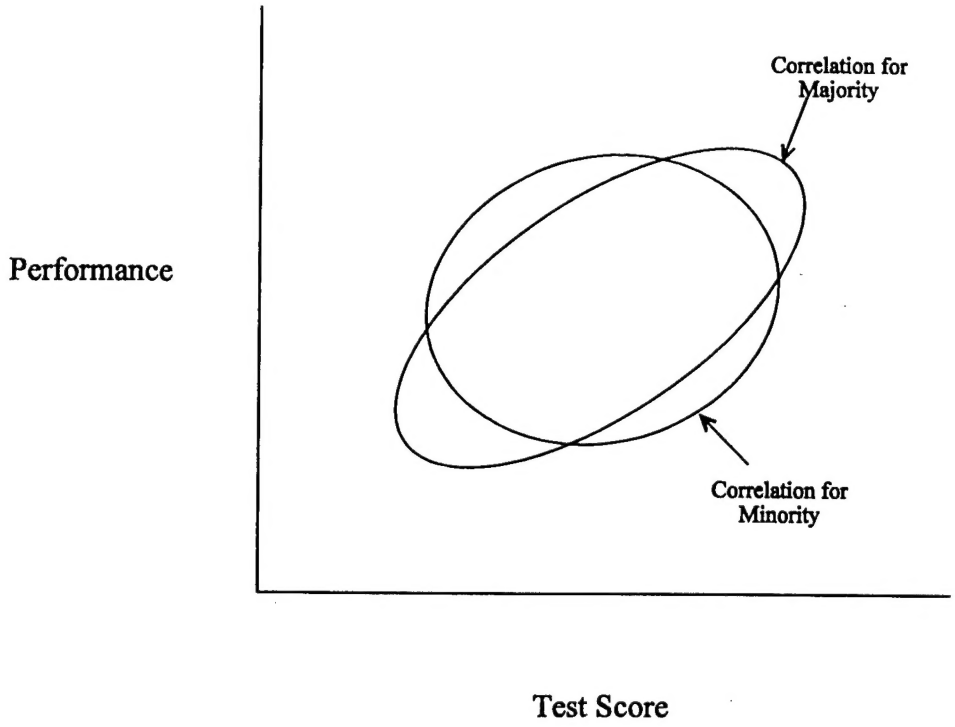


Figure 1. Differential validity: lower validity for minority than majority group

The ellipses indicate the approximate spread of individual scores from which the least-squares regression line is computed. As shown, although the validity coefficient is statistically significant for the White group, it does not achieve significance for the Black group. This would then be considered a case of differential validity and, therefore, an example of one form of test bias.

Differential prediction. While unequal validity coefficients are indicative of a biased test, equal validities do not necessarily mean that the test is unbiased according to professional consensus in the testing field. Mean differences in scores on both the predictor and criterion must also be considered.

A more comprehensive way of determining whether a test is biased is through differential prediction, defined in professional testing principles as follows:

> Predictive bias is found when mean criterion [e.g., job performance] predictions for groups differentiated on some other basis than criterion performance are systematically too high or too low relative to mean criterion performance of the groups. (SIOP, 1987, p.18)

Identification of differential prediction, then, involves the simultaneous examination of both mean differences on tests and job performance.

Figure 2 represents the "classic" form of test bias, under the differential prediction definition of bias.

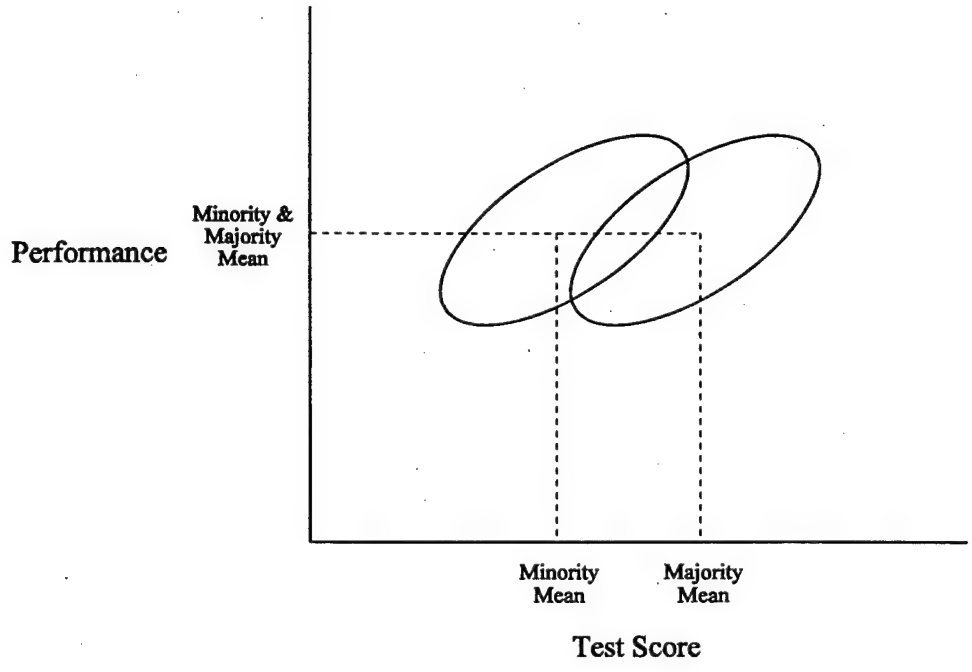


Figure 2. Test bias: mean difference on test scores but no difference in average job performance

As illustrated, there is no evidence of differential validity (as represented by the equivalent size and shape of the ellipses). There is also no difference between the average job performance of the two groups. However, there is a significant difference between the average test scores. Thus, the majority group will be hired in a larger proportion to the minority group, based on their test scores, even though both groups would perform equally well on the job. Though the example is overly simplified, the basic argument is that the mean differences between minority and majority groups in both predictor and criterion must be addressed simultaneously.

Regression lines and equations are typically used to examine differential prediction and test bias. T.A. Cleary (1968) formalized the most accepted regression-based procedure (adopted in SIOP's 1987 Principles for the Validation and Use of Personnel Selection Procedures) for determining test bias. Her model, better known as the Cleary rule, states that:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance (p.115).

Figure 3 illustrates a biased test according to the Cleary rule.

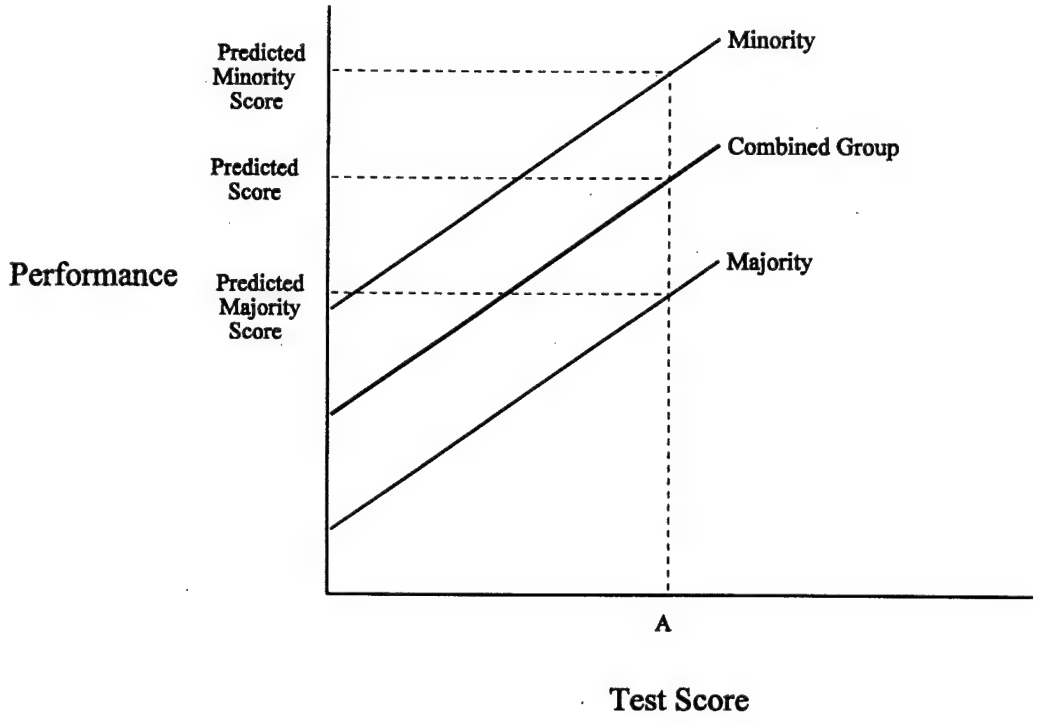


Figure 3. Common regression line overpredicts majority performance and underpredicts minority performance

This is actually the same scenario as presented in Figure 2, only using a regression line format. Included is a common regression line that is computed from the combined group (i.e., composed of both the minority and majority groups). In this instance, if a common regression line is used for both the minority and majority group, it would under-predict the job performance of the minority group, given a particular score achieved on the test (i.e., score A). It would also over-predict job performance of the majority group. Therefore, an unbiased testing procedure can only be achieved if predictions are made based on the different regression equations for the respective groups.

Thorndike (1971), however, highlighted a possible source of bias even if a test passes the Cleary rule: when the difference between the mean test scores of two groups is greater relative to the difference between their mean job performance ratings. Figure 4 illustrates this point.

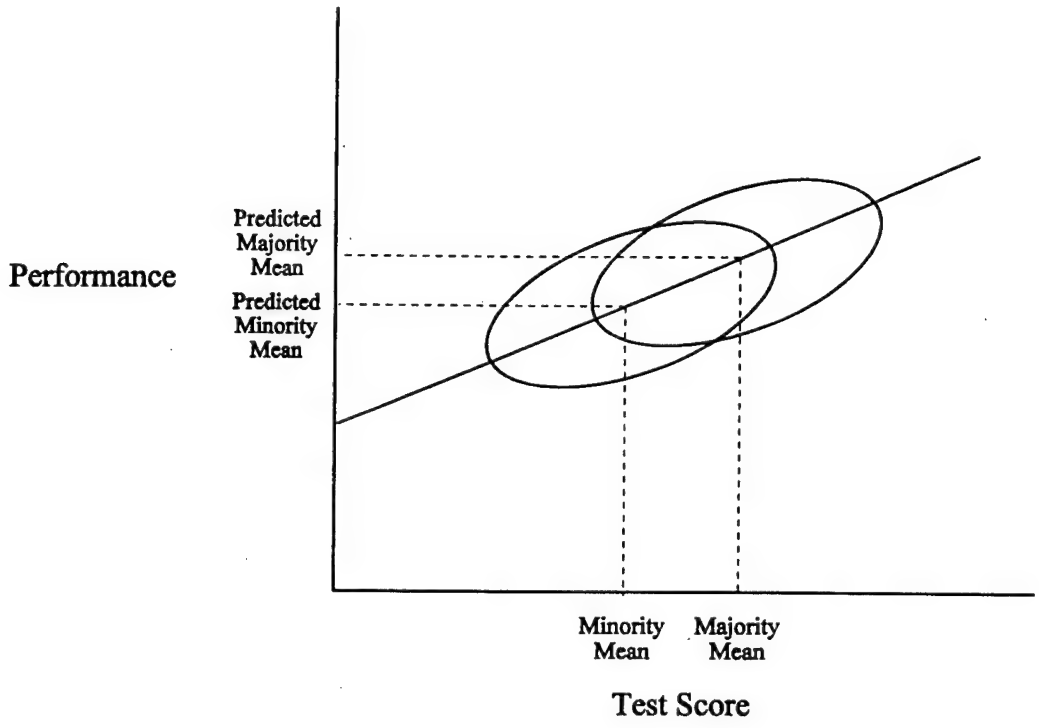


Figure 4. Average test score difference exceeds job performance difference

Assuming that both test and job performance are measured on the same scale, this hypothetical situation would be considered "fair" to both groups, according to Cleary, since they both have identical regression equations (i.e., same slope and intercept). It is considered "unfair" in the Thorndike sense because the average test score difference is larger than the average job performance difference. Thus, a larger proportion of minority

candidates would perform successfully on the job than would be suggested by their test scores.

According to Thorndike (1971), the "qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance." (p.63). Thus, a fair test in this regard, occurs only when the ratio of the proportion selected to the proportion successful is equal for both the minority and majority groups (Cole, 1973). To illustrate, if 20 percent of Group A were selected while 80 percent would have been successful performers (a ratio of 1:4) and 10 percent of Group B were selected while 40 percent would have been successful performers (a ratio of 1:4), the testing procedure would be considered fair. This is because the ratio of selected to successful job applicants is the same for both groups: 25% of applicants in both groups, who would have performed successfully on the job, are actually selected (true-positives); and 75% of applicants in both groups, who would have performed successfully on the job, are not selected (false-negatives).

Another situation where the Cleary and Thorndike models would conflict would be in their interpretation of the direction of bias. Let us assume a possible situation where the common regression line systematically over-predicted the criterion scores for the minority group and under-predicted the criterion scores for the majority group. Let us further assume that the average criterion difference is smaller than the predictor difference. Although the predictor would technically fail the test of fairness when applied against both models, the models would disagree as to the direction of bias. Cleary would

consider the bias to be working against the majority group; Thorndike would consider the bias to be working against the minority group. This is an important point to consider when evaluating not only the fairness of the predictor but also which group is being advantaged or disadvantaged by the use of the predictor.

Models of Fairness and Decision Making

The choice of which model of fairness should guide testing procedures is essentially an ideological or policy decision. Hiring on the basis of merit, also known as top-down selection, is subscribed to by many private and public sector organizations (Cascio, Outtz, Zedeck & Goldstein, 1995). This entails first selecting the applicant who scored highest on the selection test, then proceeding to the next highest scorer, an so on. Strict adherence to top-down selection, focusses on maximizing the proportion of people selected who turn out to be successful on the job (true-positives) and minimizing the proportion of people selected who are subsequently not successful (false-positives). Sackett and Wilk (1994) describe this as the perspective of the institution or organization. Of lesser concern is the problem posed by Thorndike, or the proportion of people who would have succeeded who are not selected (false-negatives).

Nevertheless, false-negatives become a central concern if workplace diversity is a valued goal. If it is, then the adoption of top-down selection is not ideal. Assuming CATs could pass the Cleary rule (this issue is dealt with is a later section), because some racial and ethnic minorities, such as Blacks, on average, score lower than Whites on CATs, they would be disproportionately burdened by false-negative evaluations. This would be true of any lower scoring group on a test with less-than-perfect prediction (see

Figure 4 and accompanying discussion).

The "conflict" between the Cleary and Thorndike models was demonstrated when the National Academy of Sciences' (NAS) panel evaluated the fairness of the General Aptitude Test Battery (GATB) developed by the United States Employment Service (see Hartigan & Wigdor, 1989). The NAS panel found little evidence of differential validity. They did, however, find that the intercept of the Black regression line fell below the White regression line, indicating that the GATB somewhat over-predicted job performance for Blacks. Thus, the test was actually somewhat biased against Whites according to the Cleary rule. But the NAS panel also found that the average Black-White gap on the job performance measure was smaller than the Black-White gap on the test. They concluded that Blacks are disproportionately burdened by a higher rate of false-negatives and, therefore, advocated score adjustment to equalize this discrepancy. They explained that:

> ...because the validities of test score for supervisor rating are modest, there is not so great a difference in average job performance between minority and majority applicants as there is in average test performance. Majority workers do comparatively better on the test than they do on the job, and so benefit from errors of false acceptance. Minority workers at a given level of job performance have much less chance of being selected than majority workers at the same level of job performance, and thus are burdened with higher false-rejection rates. (Hartigan & Wigdor, 1989, p.7).

Essentially, they adopted the Thorndike model of fairness over the Cleary model. This

decision was met with resistence and confusion on the part of the business community as demonstrated by an article in Fortune magazine. To wit:

> How can [the NAS panel] say unadjusted scores would be unfair when they just got through acknowledging that the GATB is not biased against minorities? Friends, we have looked carefully through the report's 354 pages and cannot find a crisp answer to that question. It alludes affirmatively to government policies calling for "inclusive discrimination." It reminds you, in case you forgot, that efficiency isn't everything. (More normal nonsense, p.118)

The writer appears not to understand that the Cleary model is not the only model of fairness. Nor does the model provide definitive answers to evaluate fairness (no model does). Choice of any model is dictated by one's values. Without advocating which model of fairness should be adopted, the statement reiterates a common misunderstanding of many alternative models of fairness like Thorndike's: claiming that the model ignores job performance and is nothing more than an arbitrary quota system. Job performance is taken into account but this is not always readily apparent. The NAS panel failed to make this point: although they clearly described why the GATB was not biased according to the Cleary model, they did not include in their report any mention of the Thorndike model or report the specific data showing the smaller average difference in the criterion. Consequently, they could not illustrate that the relative qualifications between minority and White workers were not as large as was suggested by the test. This oversight is remedied by the current study.

Many commentators have made a distinction between bias and fairness. Bias

refers to the invalidity of the test whereby statistical error systematically distorts the meaning of testing results for members of a particular group (Shepard, 1987). Fairness, in contrast, according to the Principles for the Validation and Use of Personnel Selection (SIOP, 1987) is a social rather than a psychometric concept. It also has no one established meaning and, consequently, lacks a single statistical or psychometric definition. Fairness, or lack thereof, is a combination of the procedure, the job, the population, and how the scores derived from the testing procedure are used. Given that it is not just reserved for use by academics and testing professionals makes it more closely tied to social policy than the concept of bias.

This study conceives of the Thorndike model as addressing this broader conceptualization of fairness (i.e., social concerns as well as accurate selection decisions) where workplace diversity as well as organizational productivity are considered. Ultimately, though, this study is an examination of CATs against one model of fairness (i.e., Thorndike) and cannot make any definitive conclusions about the fairness of using CATs in general. Such a determination can only be made by an appeal to the values of the organization, the professional guidelines it works under (if any), and the legal constraints it must work within.

Past Research: Inadequate as a Test of the Thorndike Model

It is possible to assume that the mean Black-White job performance difference (job performance disparity) will necessarily be found to be smaller than the mean Black-White CAT score difference (CAT disparity). This is because it is a statistical fact that when there is a difference between the average predictor scores of two groups, unless the

predictor correlates perfectly with the criterion, the average predicted criterion difference between groups will be smaller, assuming both groups are characterized by the same regression line. The assumption, however, is that both Blacks and Whites are characterized by the same regression line. Although it has been found that the slopes of both the Black and White regression lines are often similar, they frequently have different intercepts (e.g., Boehm, 1977; Field, Bayley & Bayley, 1977; Grant & Bray, 1970; Gael, Grant & Ritchie, 1975b; Ruch, 1972, as cited in Arvey & Faley, 1988; Schmidt, Berner & Hunter, 1973). Furthermore, the intercept of the regression line for Blacks is frequently found to be lower than the regression line for Whites, indicating that the use of a common regression line would over-predict job performance for Blacks according to the Cleary rule. Depending on the size of this intercept difference, it is quite possible that the job performance disparity could be found to be equal or even bigger than the CAT disparity. Campbell, Crooks, Mahoney, & Rock (1973, as cited in Jensen, 1980) found just that, where minorities scored about one-half SD below Whites on aptitude tests which was matched by the same difference on work samples and job knowledge tests. However, no such difference was found when supervisor ratings were used. According to Jensen (1980), this was because supervisor ratings are prone to bias whereas work samples and job knowledge are "the most objective indicators [of job performance] available." (p. 512).

What is interesting about these findings is that while differential validity of CATs has been assessed in many studies, this is not true for differential prediction. Studies used to assess differential prediction normally use a combination of ability tests, including not

only cognitive ability but knowledge and special skills such as clerical speed and accuracy (c.f., Boehm, 1977; Field, Bayley & Bayley, 1977; Grant & Bray, 1970; Gael, Grant & Ritchie, 1975b; Schmidt, Berner & Hunter, 1973). Therefore, these differential prediction studies cannot necessarily be generalized to characterize CATs exclusively. This is an easily overlooked point, as demonstrated by SIOP's (1987) assertion in the Principles for the Validation and Use of Personnel Selection Procedures that "the literature indicates that differential prediction on the basis of cognitive tests is not supported for the major ethnic groups (Schmidt, Pearlman, & Hunter, 1980; Hunter, Schmidt, & Rauschenberger, 1984)." (p. 18). Schmidt, Pearlman, & Hunter's (1980) study involved Hispanics, so is not of immediate concern to the current study. However, Hunter, Schmidt, & Rauschenberger's (1984) chapter in Perspectives on Bias in Mental Testing is again a review of the studies already mentioned that combine ability tests with other test types. Thus, direct and convincing evidence for or against differential prediction using CATs is lacking.

To sum, the Black and White regression lines have consistently been shown to have different intercepts and, therefore, the size of any job performance disparity cannot be extrapolated from the CAT disparity using a common regression line. Furthermore, the magnitude (or even the direction) of the intercept difference cannot be conclusively determined since most differential prediction studies, thus far, tend to combine other ability test measures with CATs as predictors of job performance. Thus, using the separate Black and White regression lines of the differential prediction literature to extrapolate the size of the corresponding job performance disparities would not yield

accurate results that could confidently be applied to CATs.

The purpose of reviewing the differential validity and prediction literature is not specifically intended to highlight its findings (although serious discussion should be given to the suitability of generalizing them to CATs). The review does, however, emphasize that a job performance disparity smaller than the corresponding CAT disparity is far from a foregone conclusion and has yet to be conclusively tested in the literature.

Bias in the Job Performance Measure

The discussion surrounding the fairness of CATs has largely centred on the predictor itself. Thus, little attention has been paid to the criteria that CATs are validated against. Yet, the validation of a test is dependent to a large degree on the validity of the job performance measure. Kraiger and Ford's (1985) meta-analytic investigation found that raters tended to rate people from their own race higher than those of another race. As most job performance evaluations come from supervisory ratings, coupled with the fact that supervisors are predominantly White, raises concern that past validation studies on CATs may be using biased criterion measures to the detriment of minorities. Ironically, if a biased test is validated against a job performance measure that is biased in the same direction (e.g., both biased against minorities), the result may be the conclusion that the test is fair according to the Cleary rule. There is, however, some disagreement as to the extent of rater bias. While Pulakos, White, Oppler, and Borman (1989) also found significant rater-ratee race effects, the effects accounted for less than 1% of the rating variance.

To address these concerns, CATs should be validated against equivalent objective

indices of job performance—measures that are less prone to rater bias that ideally tap the same aspects of job performance as supervisor ratings. Objective measures may include "turnover, absences, production rates, job level and salary, sales, disciplinary cases, and any other directly countable record or index." (Borman, 1991, p.301). The advantages of using objective measures of performance are that they directly record job-related behaviour, with less risk of distortion by rater bias or random error. Unfortunately, they are also considered to be very narrow indices of job performance and, therefore, not as complete as supervisory ratings. This is a reasonable concern which is addressed by Nathan & Alexander (1988) and Hoffman, Nathan & Holden (1991) who found that subjective and objective measures were both predicted well by measures of cognitive ability, lending support for the equivalency of the two types of job performance measures. Martocchio and Whitener (1992) rightly point out, though, that these studies follow the "differential validity" paradigm, speaking only to the slopes of the objective and subjective measures' respective regression lines, not their intercepts. Furthermore, Bommer, Johnson, Rich, Podsakoff & Mackenzie (1995) found that objective and subjective measures were not correlated highly enough with each other to be considered interchangeable. The various studies are not so much contradictory as they are incomplete concerning the equivalency of objective and subjective measures of job performance. Therefore, no conclusive evidence exists to establish the valuing of one type of job performance measure over another.

Finally, Ford, Kraiger and Schechtman (1986) investigated the impact of using objective versus subjective measures of job performance when evaluating Blacks and

Whites. They found that average performance differences between Blacks and Whites were greater (in favour of Whites) when subjective measures were used compared to objective measures. This is further evidence to suggest that subjective ratings may indeed be subject to rater bias to the detriment of Blacks.

Studies Investigating the Comparability of CAT and Job Performance Differences

Two previous studies have addressed the bias issue in a manner similar to the methodology used in this thesis. Schmitt, Clause and Pulakos (1996) analysed average score differences between Blacks and Whites on CATs and job performance measures. They found a .83 SD difference between groups on CAT scores in favour of Whites, somewhat smaller than the 1 SD commonly cited in the literature. Job performance measures typically predicted by CATs had markedly smaller differences between Blacks and Whites: .15 SD in clerical speed/accuracy, .33 SD in accomplishment record, .38 SD in job sample/job knowledge. This is certainly suggestive of bias when CATs are used for personnel selection, with fairness by the Thorndike definition decreasing when objective measures are used, such as clerical speed and accuracy. Unfortunately, the number of effect sizes was relatively small across measures. Also, the researchers did not require that studies included in their meta-analysis have a matched sample of a cognitive ability predictor and a job performance measure for the same job. Such a restriction could introduce sample-specific differences in underlying ability that could differentially affect CAT and job performance scores.

Martocchio and Whitener's (1992) study extends the results of Schmitt et al's (1996) (although their study was published prior to Schmitt et al.'s) by using a matched

sample of CATs and job performance measures (both objective and subjective). They found evidence for the unfair use of CATs when they studied average White and non-White score differences. The mean difference between CAT scores was .46 SD in favour of Whites, again, much smaller than 1 SD. Subjective criteria resulted in a .28 SD difference in job performance between groups and objective criteria resulted in a -.009 SD difference in favour of non-Whites. These findings would suggest that not only are job performance differences much smaller than CAT differences, but when using the job performance measures that are less susceptible to bias, no meaningful performance difference between Whites and non-Whites is observed. Given that the subjective and objective measures assessed the exact same performance dimensions eliminates the concern over comparing nonequivalent aspects of job performance. Unfortunately, the number of studies used was relatively small (only eight), resulting in less than 25 effect sizes across measures. Minorities were also collapsed into the same category, which conflates the performance of Blacks with those of other minorities, including Asians and Hispanics. Finally, only one study was published after 1980 and, therefore, the bulk of the data is at least 20 years old. As will be elaborated on in the next section, mean CAT score differences between races have been shown to be shrinking over time, and therefore Martocchio and Whitener's study may already be outdated in that it does not reflect the current US population.

To sum, the current study is an improvement over these past studies offering a test of the Thorndike model of test fairness for the following reasons:

1. The number of studies and effect sizes is significantly larger.

2. The study uses a matched sample, i.e., a CAT score and job performance rating were available for each subject.

3. The minority groups are not combined. Only Blacks and Whites are compared.

4. The data are much more recent. Prior meta-analyses (Martocchio & Whitener, 1992; Schmitt, Clause and Pulakos, 1996) incorporated studies that were mostly conducted prior to the 1980s.

Despite the deficiencies of previous studies, they are nonetheless suggestive that CATs would not meet the conditions of the Thorndike model. The current meta-analysis, by addressing these deficiencies, is intended to draw more definitive conclusions about whether CATs accurately reflect the relative job qualifications between Blacks and Whites and what the results mean for Black representation in the workplace.

Summary

- A common regression line cannot be assumed. There is ample evidence to suggest that the intercepts between Black and White regression lines are different. As a result, there is no statistical reason why job performance differences might not be equal to or greater than CAT differences.
- Thorndike has shown, conceptually, that a test that passes the Cleary rule can still be considered unfair. Assessing the job performance disparity against the CAT disparity under the Thorndike model is the main goal of this study. This focus is highly relevant to Employment Equity (E.E.) concerns since it demonstrates the eventual proportion of minorities qualified to do the job (if any) who are denied entry based on CAT scores. It deserves a more definitive test in its own right.

Scientific Racism and the Interpretation of Mean Differences

Given that racial research has been misused to advance racist causes (see Tucker,1994, for a detailed review) this issue warrants a special discussion that places this study (and similar ones that compare mean differences between groups) in a wider context. The main concern is that unwarranted conclusions may be drawn from this study's findings. Specifically, the average score of American Blacks on CATs are often reported to be lower than the average score of American Whites (Jensen, 1980). It is not clear why Blacks and Whites in the US, on average, differ in their CAT scores. Neisser et al. (1996) and Frisby (1995) discuss possibilities ranging from lower mean income; inadequate schools; cultural differences; and low self-efficacy, self-esteem, and achievement motivation due to discrimination. Since the current study does not control for these influences, broader conclusions as to why there are average differences between Blacks and Whites cannot be determined from the data collected.

The fairly robust finding of an average one SD difference between American Blacks and Whites on CAT scores from the 1930s to the 1980s may imply that very little has changed in five decades (Neisser, 1998). However, this ignores the observation that CAT scores across races are rising. The average IQ score of Black Americans in the 1980s is roughly the same as those of White Americans in the 1930s (Neisser, 1998). This phenomenon has been dubbed the "Flynn effect", named for James Flynn who systematically documented the score increases over time (Flynn, 1984, 1987, 1999). Thus, mean differences cannot be considered absolute or immutable values characterizing the intelligence of either Blacks or Whites. Moreover, the Black-White gap in CAT

scores may not be as enduring as once thought. For example, in 1978, the Black-White difference in the math scores of 17-year-olds was about 1.1 SD. By 1990, the difference was about 0.6 SD. Similar trends were found for verbal scores (Neisser, 1998).

The preceding discussion of the Black-White gap in average CAT scores should, however, not be taken to mean that CAT scores can be considered in isolation of the criteria they are meant to predict. Scores on intelligence tests have been inappropriately reified by researchers such as Jensen (1980) and used to rank people and races in order of comparative worth—worth not based on scientific analysis but determined by racist ideology (see Tucker, 1994, for an historical discussion of the advancement of racism through the use of intelligence tests). But, as noted by Campbell (1996):

> ...mean differences are, or should be, of no intrinsic interest. Their importance derives exclusively from the value of changes in the dependent variables that cognitive abilities predict. For example, if IQ was not related to anything deemed important then IQ differences between people or between groups would be of no interest. (p. 133).

Thus, for the purposes of the current study, mean differences observed between groups on CATs should only be considered in light of the differences (if any) found in average job performance.

Also of concern is that the mere act of analyzing people by race may perpetuate the misconception that race, in and of itself, is responsible for any observed differences between groups. This may not only overlook the aforementioned differences in the environmental and social experiences of racial groups, but it may also ignore the within-

group variation of different groups. Categorizing people by race treats groups as if they were a relatively homogenous collection of people, when in fact, they are not. The variation within groups is actually much greater than the variation between groups. Thus, differences found between different races are actually smaller than differences found between people of the same race.

The consequences and misuse of racial research has a long history and has been the topic of much controversy. It is beyond the scope of this paper to discuss these issues in any detail and the interested reader is referred to Tucker (1994) and Winston (1996, 1998) for fuller discussions. This brief overview is intended to caution the reader that the reasons as to why there are average racial differences in CAT scores are not clear and cannot be inferred from the findings of this study. Nor can CAT scores be considered in isolation of what they are designed to predict. Interpretations of this study should be confined to the comparison of CAT scores and job performance measures.

Method

Overview

The principal aim of this study is to evaluate whether CATs meet the conditions of a fair test according to the Thorndike model: that is, evaluating whether the standardized mean difference between Black and White CAT scores (racial CAT disparity) is reflected by a corresponding and equal mean job performance difference (racial job performance disparity).

Secondly, it will be determined if objective and subjective measures of job performance result in equal racial job performance disparities. This is an important test

since, thus far, CATs have primarily been validated against subjective measures such as supervisor ratings which, it was argued in the introduction, are more prone to rater bias. If subjective and objective measures result in different racial job performance disparities, this might indicate that subjective measures are being systematically influenced by rater bias and, in turn, bias the results of differential prediction studies (i.e., determining bias against the Cleary rule) that rely on subjective measures of job performance.

Thirdly, if CATs do not meet the conditions of a fair test according to the Thorndike model, the degree to which CATs misrepresent the relative qualifications between Blacks and Whites will be evaluated. This will include determining the extent of false-negatives for Blacks over and above those that occur for Whites that would result given various test cutoff scores.

Sample

PsycInfo, ERIC, Wilson Business Abstracts, ABI/Inform (Business), Dissertation Abstracts, and the Annual Review of Psychology, were reviewed, in addition to contact with several government, military, academic, and private organizations, to identify published and unpublished studies for inclusion in the meta-analytic database. Selected studies were required to have: 1) a written cognitive ability/intelligence test as well as at least one measure of job performance, 2) means and standard deviations of the CATs and job performance measures categorized by race (i.e., Black and White), and 3) no less than 10 subjects in each racial subgroup. Using these sources, and applying the above three criteria, 39 studies were provisionally deemed suitable for the meta-analysis. However, about two-thirds of the studies were found to have insufficient data and the researchers

were contacted to provide the required data. This is indicative of a trend away from reporting means and standard deviations by subgroup since the beginning of the 1980s. In the final result, 20 studies provided the necessary data for analysis.

The General Aptitude Test Battery (GATB) validation database of the U.S. Department of Labor, provided by the National Center for O*NET Development, contributed 115 additional studies, resulting in total of 135 studies.

The GATB and non-GATB data were analysed separately as well as together for the following reasons:

- It was not possible to determine the quality of the studies used in the GATB database and, given its large size relative to the remaining studies in the database for this study, any systematic problems with the data would unduly impact the results of the meta-analysis.
- Use of the GATB might distort the comparison of subjective and objective measures in as much as the GATB database only makes use of subjective ratings.
- Because the GATB data used a single measure of cognitive ability and the same method for evaluating job performance across all its studies (i.e., an overall job performance rating scale), its separate analysis provided an opportunity to determine the extent to which the type of predictor and criterion measures were moderating influences on the results of the meta-analysis.

Coding

The researcher coded the above studies into the meta-analysis database using the following conventions.

Cognitive ability tests. CATs were considered, at a minimum, to be composed of a numerical or verbal component. Where separate subtests of verbal and numerical aptitude were reported without a composite score, these subscores were averaged together. If the sample sizes differed between the numerical subtest and verbal subtest, they were averaged in the calculation of the pooled standard deviation. Where more than one CAT was used, their scores were averaged into a single composite score. Only the general learning ability or intelligence subscore of the GATB was used to represent CATs because the GATB is also made up of nonintellective factors such as manual dexterity. 34 effect sizes were derived from the accumulated, non-GATB studies and 115 effect sizes came from the GATB database.

Job performance. Job performance was considered any evaluation of either overall job competence, particular aspects of job performance, or results from satisfactory or unsatisfactory job performance such as awards and promotion. Work samples such as assessment centres were also considered to be reflective of job performance (Borman, 1991). Job knowledge, paper-and-pencil tests were not included, because, in the researcher's opinion, they resembled a testing situation closer to CAT conditions than job performance conditions. This resulted in the loss of three studies. Job performance ratings were coded as either subjective or objective. Subjective measures were composed of supervisor and instructor evaluations. Objective measures included turnover, absenteeism, speed, accuracy, accidents, etc. (see Appendix for a comprehensive listing). Where more than one of the same type of job performance rating (either subjective or objective) was taken within the same performance domain or a global performance rating

was given for two or more performance dimensions, the scores were averaged to form a composite measure of job performance. If sample sizes differed between the performance measures being averaged, then the sample sizes were averaged when computing the pooled standard deviation. 172 effect sizes for subjective measures (57 non-GATB and 115 GATB) and 30 for objective measures (all non-GATB) were derived from the accumulated studies.

The Appendix summarizes the studies included in the meta-analysis. Listed are author(s), number of subjects separated by race, individual effect sizes, type of jobs, and cognitive ability and performance measures.

Analysis

The effect size d was first computed for the cognitive ability and job performance measures in each study—d being the mean score of the White group minus the mean of the Black group divided by the pooled standard deviation. This standardized difference score, however, has a small sample bias. The meta-analytic program by Schwarzer (1991), used in this study, corrects for this bias according to the correction procedures outlined by Hedges and Olkin (1985). Although Hedges and Olkin refers to d as g, this study will adopt the more common convention of referring to the biased effect size estimator as d (in part, this is to avoid confusion with the use of g as representing general intelligence). For every meta-analysis conducted, the weighted, unbiased mean effect size estimator d_+ was computed.

To determine whether the individual effect sizes were consistent across studies (i.e., sharing a common effect size), the homogeneity statistic Q and percentage of

variance attributable to sampling error were consulted. Heterogeneity of the data indicates the presence of moderating variables (Hunter & Schmidt, 1990; Schwarzer, 1989). Because the GATB data used a single measure of cognitive ability and a consistent method for evaluating job performance across all its studies, its separate analysis provided an opportunity to determine the extent to which the type of measures were moderating influences on the results of the meta-analysis.

Additional Meta-Analytic Considerations

Given that meta-analytic research requires many judgement calls when coding data, the following is a review of the decisions made and rationales behind those decisions.

- Studies were required to contain both a CAT and a job performance measure and, thus, the present study is considered to be composed of entirely matched samples. However, in Martocchio and Whitener's (1992) meta-analytic study, the matched sample contained a CAT, a subjective criterion measure and an objective criterion measure. This, unfortunately, restricted their sample considerably since few validation studies use more than one criterion measure. This thesis included a study where there was a CAT reported and either an objective or subjective job performance measure.
- Although it is possible to adjust for the unreliability of the measures, the required information to do so was not always available in the studies used. Therefore, adjustments for measurement error were not made on either predictor or criterion.
- Neither predictor nor criterion were corrected for range restriction. Although

restriction of range may underestimate the size of the standardized difference scores across measures, it should not impact the comparative size of the differences between measures, making this adjustment unnecessary.

- Individual effect sizes ($\underline{d}$) are assumed to be independent of each other (Hunter & Schmidt, 1990). Thus, only one effect size should come from each study. When a study has multiple effect sizes, they should be averaged together. This was done when the multiple effect sizes were measuring the same construct (see coding section). However, more than one effect size was taken from some studies when they measured different types of job performance. As long as the number of effect sizes contributed by one study are few relative to the total number of effect sizes, the error in the resulting cumulation is small (Hunter & Schmidt, 1990). Furthermore, while violations of the independence assumption affect (inflate) the observed variance, they have no systematic effect on the average $\underline{d}$. Again, violations of assumptions must be weighed against the potentially greater impact of losing data.
- One study by Roberts and Skinner (1996) contributed 2 effect sizes with subgroup sample sizes of 12, 453 each. The analysis was rerun without these effect sizes with no appreciable difference in the value of $\underline{d}_+$. Thus, Roberts and Skinner study was included in the final meta-analysis.

Results

Table 1 summarizes the results of the meta-analysis. The $\underline{d}_+$ scores listed in the third column are the average standardized difference scores or racial score disparities

between Blacks and Whites.

Table 1
Meta-analytic results

Differences between Blacks and Whites using:	k	d_+	95% confidence interval for d_+	% var. due to sampling error	Q
Non-GATB data					
CAT	34	0.68	0.57 to 0.79	27.56	155.89*
Job Performance	87	0.24	0.16 to 0.31	24.54	515.38*
Subjective	57	0.30	0.22 to 0.39	27.46	340.10*
Objective	30	0.12	0.00 to 0.24	24.48	155.45*
GATB data					
CAT	115	1.12	1.06 to 1.17	44.67	245.08*
Job Performance (subjective)	115	0.38	0.34 to 0.43	56.63	214.52*
Non-GATB + GATB data					
CAT	149	1.01	0.96 to 1.07	29.97	706.42*
Job Performance (subjective & objective)	202	0.32	0.28 to 0.36	35.56	803.62*

Note. k = number of effect sizes; d_+ = average unbiased effect size estimator (Hedges & Olkin, 1985); Q = homogeneity statistic.
*$p < .0001$.

As shown for the measures in the total sample (see Non-GATB + GATB data in Table 1), while Blacks, on average, score about 1 SD lower than Whites on CATs, the actual job performance difference between Blacks and Whites is considerably less. The difference between Blacks and Whites in their average job performance is only 1/3rd that of their average CAT difference.

Without the GATB data (see Non-GATB data in Table 1), the Black-White gap in CAT scores shrinks markedly, from 1.01 to 0.68. Since the non-GATB data consists of more recent studies, this smaller racial disparity possibly supports previous observations that the Black-White gap in CAT scores is shrinking. In terms of the CAT-job

performance comparison, the non-GATB data reveal that the job performance disparity is again, 1/3rd the size of the CAT disparity. This is due to a similar reduction in the job performance disparity. Figure 5 graphically represents the relative qualifications between Blacks and Whites as expressed by CATs and as reflected in actual job performance. Note the discrepancy between the CAT difference and the job performance difference between Blacks and Whites.

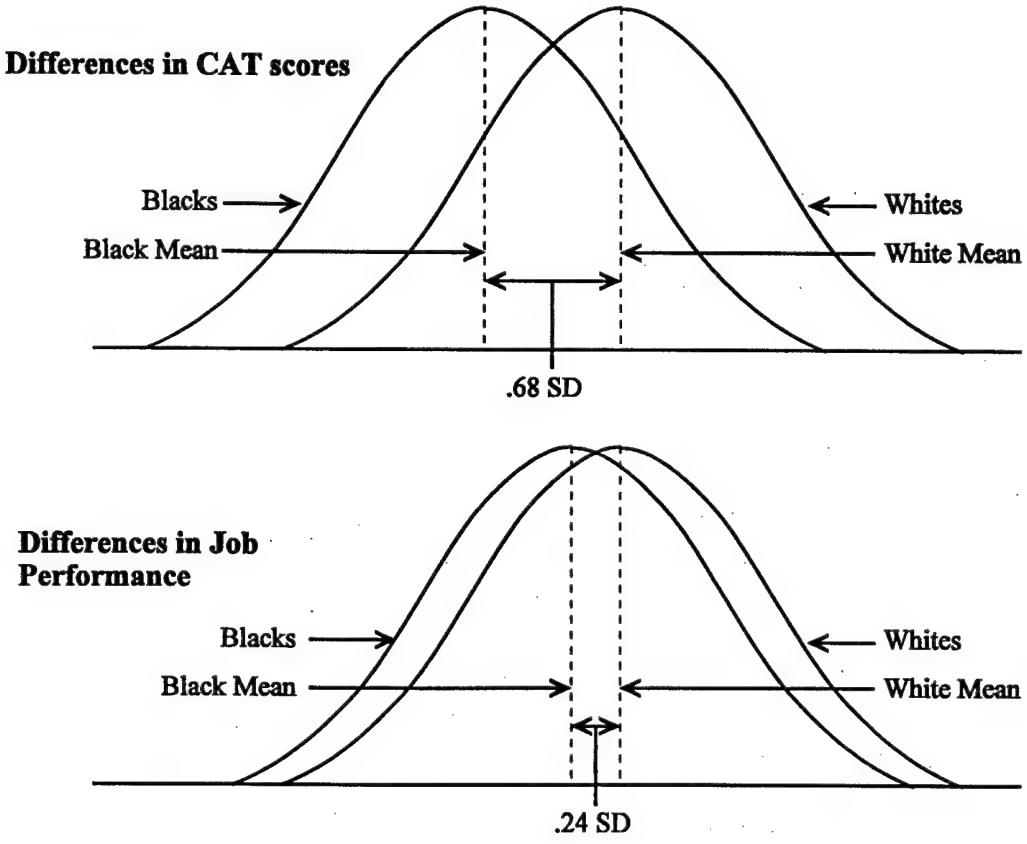


Figure 5. CAT and job performance differences between Blacks and Whites

Thus, CATs tend to exaggerate the difference between Blacks and Whites in their relative

qualifications, with CAT scores predicting the average job performance difference between groups three times larger than is actually the case.

Using the non-GATB data only, job performance was separated into subjective and objective categories. Objective measures—job performance indices less prone to rater bias—were found to have a racial disparity less than half the size of the racial disparity found using subjective measures. This may indicate that subjective measures are indeed being systematically affected by rater bias to the detriment of Blacks.

The hypothesis of homogeneity was rejected in every case using the overall fit statistic calculated by Hedges and Olkin's (1985) procedure (see Q statistics of Table 1). Significant results suggest the presence of moderating variables. Furthermore, the amount of variance accounted for by sampling error was small, less than 28% in all cases for the non-GATB data, far less than the 75% minimum suggested by Hunter and Schmidt (1990) as adequate to rule out the influence of moderators. Sampling error did account for more variance when only the GATB data was analysed. This suggests that the type of measures used to evaluate both CAT and job performance may moderate the extent of mean differences, although other moderators (e.g., job type) may be at work as well.

Discussion

The results indicate that cognitive ability tests (CATs) substantially misrepresent the relative qualifications that exist between Blacks and Whites: the actual average job performance difference between Blacks and Whites being 2/3rds smaller than is predicted by CATs, with an even greater disparity when objective measures of job performance are

considered. Furthermore, if we accept the argument that objective measures are less prone to rater bias than subjective measures, then the larger racial disparity found on the subjective measures may indicate that rater bias is affecting subjective job performance evaluations to the detriment of Blacks. Given that most validation and differential prediction studies using CATs rely on subjective ratings of job performance, care must be exercised in interpreting these studies as tests of the Thorndike model—or any other model—of test fairness.

The results of this study have considerable implications for the use of CATs in personnel selection. The following applies the CAT and job performance differences to a typical selection scenario. If one assumes a 50% selection rate for the White applicant group, based on the 0.68 SD difference observed between Blacks and Whites on CATs, only 24% of Blacks would be selected (i.e., considered qualified) using a top-down selection process. However, based on the actual job performance difference of 0.24 SD, about 40% of Blacks should have been selected (i.e., are considered qualified). Figure 6 is the same as figure 5 that shows the relative qualifications between Blacks and Whites, but also includes the cutoff score presented in this scenario.

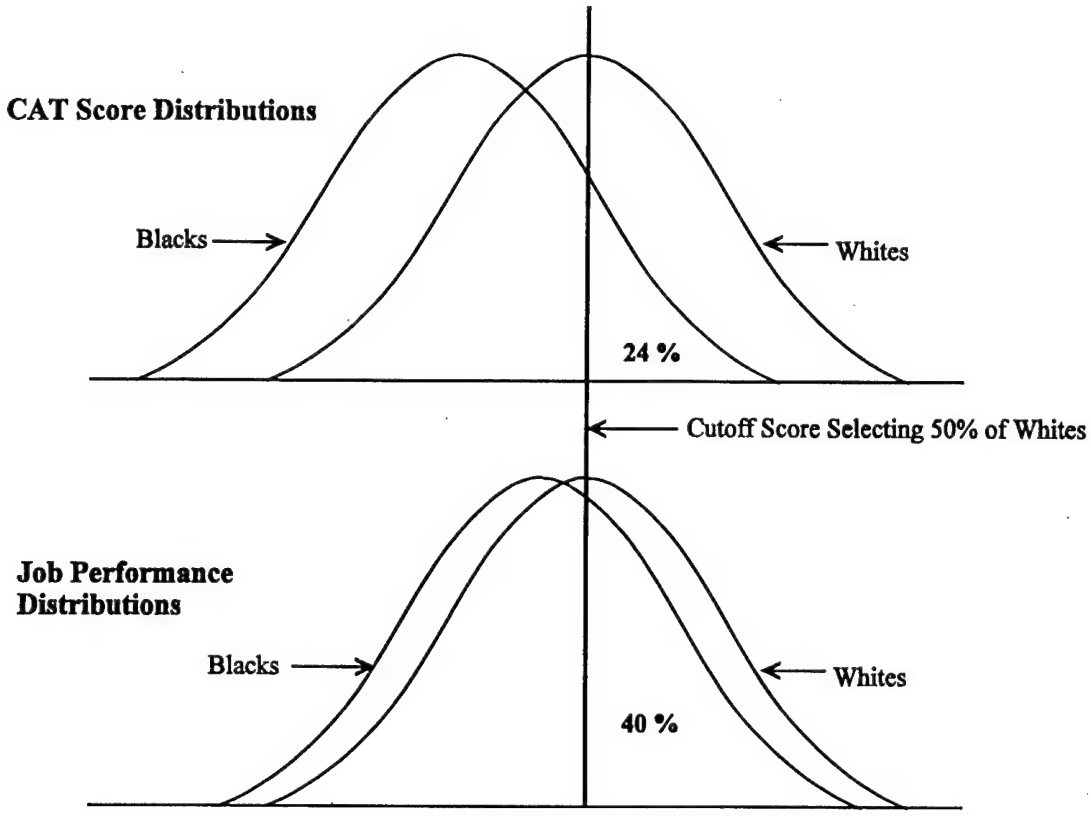


Figure 6. Proportion of Blacks selected based on predicted and actual job performance

The selection rate of Blacks should have been still higher (with an observed .12 SD difference) if objective criteria were used as measures of job performance. Thus, in the above scenario, about 40%-48% (using subjective and objective job performance measures respectively) of Black workers who should have been selected for the job, would have been incorrectly eliminated based on CAT scores. What is worse, differences between Whites and Blacks in false-negative selection rates becomes more pronounced as test cutoff scores are set higher. At a more realistic 10% selection rate for Whites, a

staggering 62%-71% of Blacks would be incorrectly eliminated. Table 2 is a complete comparison of selection ratios given particular standardized group differences.

Table 2
Minority group selection ratios when the majority group selection ratio is .01, .05, .10, .25, .50, .90, .95, or .99

Standardized group difference (d)	Majority group selection ratio								
	.01	.05	.10	.25	.50	.75	.90	.95	.99
0.0	.010	.050	.100	.250	.500	.750	.900	.950	.990
0.1	.008	.041	.084	.221	.460	.716	.881	.938	.987
0.2	.006	.033	.069	.192	.421	.681	.860	.925	.983
0.3	.004	.026	.057	.166	.382	.644	.837	.910	.978
0.4	.003	.021	.046	.142	.345	.606	.811	.893	.973
0.5	.002	.016	.038	.121	.309	.568	.782	.873	.966
0.6	.002	.013	.030	.102	.274	.528	.752	.851	.957
0.7	.001	.010	.024	.085	.242	.488	.719	.826	.947
0.8	.001	.007	.019	.071	.212	.448	.684	.800	.936
0.9	.001	.006	.015	.058	.184	.409	.648	.770	.922
1.0	.000	.004	.011	.047	.159	.371	.610	.739	.907
1.1	.000	.003	.009	.038	.136	.334	.571	.705	.889
1.2	.000	.002	.007	.031	.115	.298	.532	.670	.869
1.3	.000	.002	.005	.024	.097	.264	.492	.633	.846
1.4	.000	.001	.004	.019	.081	.233	.452	.595	.821
1.5	.000	.001	.003	.015	.067	.203	.413	.556	.794

Note. From "The Effects of Forming Multi-Predictor Composites on Group Differences and Adverse Impact," by P. R. Sackett and J. E. Ellingson, 1997, Personnel Psychology, 50, p.710.

Thus, the use of CATs in personnel selection under realistic selection scenarios will result in a large underrepresentation of Blacks in the workplace. From an E.E. perspective, this is not justifiable because the pool of qualified Black candidates, relative to Whites, is considerably larger than is suggested by CAT scores.

Limitations

This study was grounded on many of the assumptions that past validity

generalization studies of personnel selection measures have been based (e.g., U.S. Department of Labor, 1983). These assumptions centre on the accuracy of the job performance measures themselves. While comparison of the objective versus subjective results in this study are indeed suggestive that supervisor and instructor ratings of job performance may be biased, caution should be used when interpreting any job performance measure and subsequently, the tools validated against them. It is acknowledged that job performance is, in fact, multidimensional in nature, involving many performance components (Campbell, 1990). The fact that these various performance components may not correlate well is problematic when employing a composite measure of performance, as is often the practice in performance evaluations. If a candidate performs well on one measure and poorly on another, a combining of the two results in the attenuation of both (Guion, 1998).

Another problem, as already discussed, is that objective and subjective performance measures may not be equivalent in terms of the aspects of performance they measure. Objective indices are often considered incomplete measures of job performance. While a case may be made for the non-equivalency of objective and subjective measures of job performance, it should not be assumed that subjective ratings are superior indicators of job performance simply because they combine more performance indicators into a single score. Not only does a single score have its weaknesses, as already mentioned, but what exactly is being measured is sometimes obscured and lacks precision. The Standard Descriptive Rating Scale used as the criterion in the GATB studies included in this research, is a prime example, which is

generic in nature and not occupation specific. Furthermore, the fact that organizations have chosen to track specific performance measures that can be classified as objective in nature, implies that they are essential aspects of the job. Again, what constitutes a superior job performance measure has not yet been resolved.

The objective and subjective categorizations also may not be as distinct or mutually exclusive as perhaps is suggested. For example, the standards of job performance, as measured by objective criteria, are based on a subjective decision (Nathan & Alexander, 1988). Thus, categorizing a measure as "objective" may hide the "subjective" decisions that went into formulating the measure. Nonetheless, we can be relatively confident that the opportunity for bias entering performance evaluations are less for objective measures than they are for subjective measures, though not completely removed.

The moderating influences of other factors on observed variance could not entirely be accounted for by sampling error or the type of predictor and criterion measure used. Other factors (perhaps including type of job or organization) contribute to differences between studies in the extent of CAT and job performance disparities. Future research may help to further refine our knowledge of these moderator variables.

Finally, the results of this study should not be used to make any inferences or generalizations about Blacks or Whites as a whole. The samples used consisted predominantly of populations residing in the United States and, therefore, cannot be generalized to populations from other countries. Furthermore, the data analysed were, for the most part, from job incumbents, which is a specific subpopulation. The actual

qualifications and backgrounds of these people are unknown and cannot be assumed to be indicative of more general populations. Currently, our understanding of how representative the samples of job incumbents are of their respective groups is limited. For example, Blacks and Whites, in all likelihood, are impacted differently by such diverse factors as socioeconomic status; barriers to education and other resources; self-selection; recruiting tactics of the organizations; systemic racism, etc. How and to what extent these and other factors influence the numbers of Blacks and Whites who apply for specific jobs requires further investigation. Thus, applying the results of this study to make characterizations about any racial groups in general would be inappropriate (refer back to the section on Scientific Racism).

Implications

The results of this study show that CATs fail the Thorndike test of fairness. Thorndike himself advocates adjusting CAT scores to offset the incidences of false-negatives that disproportionately burden Blacks. However, Thorndike's method and other models of fairness have been criticized as quota setting as well as on grounds that they do not maximize the utility of the testing procedure. As the Civil Rights Act of 1991 has banned any form of score adjustment based on "race, color, religion, sex, or national origin" (Pub. L. No. 102-166, Section 106) the controversy surrounding score adjustments has effectively been rendered moot. I will, therefore, not pursue this line of redressing test unfairness against Blacks. The methods and merits of score adjustments can be found in Darlington (1971), Thorndike (1971), Cole (1973) and Hunter, Schmidt and Rauschenberger (1984). Perhaps CATs should not be used for selection under any

circumstances (score adjusted or not).

One point, however, does deserve comment. The characterization of score adjustments as quota setting can be misleading (I make no claim as to the intentions of the authors who make such characterizations). Whether deservedly or not, the term "quota" has been stigmatized to mean an arbitrary decision to increase minority representation equal to the representation of the majority group, regardless of qualifications or ability. This is a fundamentally inaccurate characterization of the Thorndike test fairness model. Score adjustments are based on realized job performance and would not result in equal representation unless there are no average differences between groups in job performance. But, to use this study as an example, the use of CATs does substantially misrepresent the relative qualifications between Blacks and Whites. Therefore, score adjustments (if used) should align the testing procedure with the actual level of job performance achieved by both groups.

A final determination of whether to use CATs—and in what capacity—cannot be addressed by this study alone. As noted previously, it is a function of both the psychometric properties of the test (i.e., validity & reliability) as well as societal concerns, the values of the organization, and the professional and legal guidelines to which the organization must adhere. The results of this study do, however, provide additional fairness information on CATs that can supplement the findings of differential prediction studies that adhere to the Cleary interpretation of fairness. I also believe that this study more directly addresses E.E. concerns in that it not only speaks to the barriers that CATs present to workplace diversity, but it also illustrates that the relative

qualifications between Blacks and Whites are not as different as CATs would predict.

Future Research

Helms (1997) states that the cultural equivalence of CATs for different ethnic and minority groups has not been addressed adequately by test developers. The racial, ethnic culture, and socioeconomic conditions of socialization are rarely applied, integrated into or removed from CATs. In part, this is because these concepts are poorly operationalized or understood by test developers. Nevertheless, some information suggests that these domains of socialization uniquely contribute to CAT performance (e.g., Grubb & Dozier, 1989; Robinson, 1994, 1995; as cited in Helms, 1997). Helms (1997) suggests a broader conceptualization of cultural equivalence that moves beyond the simple removal of culture specific language. Among the lesser known forms of cultural equivalence, she identifies:

> (e) [sic] testing condition equivalence, assurance that the idea of testing as a means of assessing ability and the testing procedures are equally familiar and acceptable to Blacks (and other [visible racial/ethnic groups]) and Whites . . . and (g) sampling equivalence, determination that samples of subjects representing each racial or ethnic (or cultural or socioeconomic) group are comparable at test development, validation, and interpretation stages. (Helms, 1992, p.1092; as cited in Helms, 1997)

Future research should be applied to developing CATs that are mindful of these concerns (see Helms, 1997, for a comprehensive discussion) and subsequently validated against relevant job performance dimensions.

Regardless of the resolution to the CAT fairness/bias question, it will still not resolve the problem of the role of nonintellective[2] factors and adaptive skills that may contribute to successful job performance such as motivation. Efforts should be continued to identify viable alternatives to CATs—comparable in efficiency and validity—that result in less adverse impact. Methods such as training and experience ratings already yield validity coefficients comparable to that of CATs (McDaniel, Schmidt & Hunter, 1988). There is, however, a need for research evaluating the fairness of alternative measures with regards to both differential prediction and mean difference studies.

Finally, as observed by previous meta-analysts, there has been a decline in the documentation of means and standard deviations throughout the published literature. This state of affairs often limits the scope of meta-analytic investigations—that ask similar questions to this one—by restricting the number of representative studies that can be included. An appeal is made to reestablish the importance of publishing descriptive statistics.

Summary and Conclusion

The results of this study found that CATs substantially misrepresent the relative qualifications between Blacks and Whites. Moreover, the extent of this misrepresentation is more pronounced when less biased measures of job performance are used. Under realistic selection scenarios, these findings would indicate that the use of CATs would result in the underrepresentation of Blacks in the workplace. From an E.E. perspective,

---

[2]I would qualify this statement by contending that nonintellective factors would, in fact, include aspects of intelligence, as of yet, unmeasured by current CATs, in addition to other factors.

this would not be justifiable because the pool of qualified Black candidates, relative to Whites, is considerably larger than is indicated by the use of CAT scores. Finally, it is hoped that this study provides an increased appreciation of the relevance of the Thorndike model as an additional consideration when evaluating the fairness of CATs or other tests for personnel selection.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Arvey, R. D. & Faley, R. H. (1988). Fairness in Selecting Employees (2nd ed.). Don Mills, Ontario: Addison-Wesley Publishing Company.

*Baehr, M. E., Saunders, D. R., Froemel, E. C., & Furcon, J. E. (1971). The prediction of performance for Black and for White police patrolmen. Professional Psychology, 2, 46-57.

Boehm, V. R. (1977). Differential prediction: A methodological artifact? Journal of Applied Psychology, 67, 146-154.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & Mackenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. Personnel Psychology, 48, 587-605.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of Industrial and Organizational Psychology (2nd ed., Vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists Press, Inc.

Campbell, J. P. (1990). Modelling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of Industrial and Organizational Psychology (2nd ed., Vol. 1, pp. 687-732). Palo Alto, CA: Consulting Psychologists Press, Inc.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. Journal of Vocational Behavior, 49, 122-158.

Casio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1995). Statistical

implications of six methods of test score use in personnel selection. Human Performance, 8 (3), 133-164.

Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (Nov. 21, 1991).

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cole, N. S. (1973). Bias in selection. Journal of Educational Measurement, 10 (4), 237-255.

*Dalldorf, M. R. & Holmgren, R. L. (1993). A validation of the ASVAB against supervisors' ratings in civilian occupations (Contract No. MDA 903-89-C-0255). Palo Alto, CA: American Institutes for Research.

Darlington, R. B. (1971). Another look at "culture fairness." Journal of Educational Measurement, 8 (2), 71-82.

*DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Folgi, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and White-Black differences. Journal of Applied Psychology, 78(2), 205-211.

*Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as a moderator of the prediction of job performance. Personnel Psychology, 24, 609-636.

Field, H. S., Bayley, G. A., & Bayley, S. M. (1977). Employment test validation for minority and nonminority production workers. Personnel Psychology, 30, 37-46.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains. Psychological Bulletin, 95, 29-51.

Flynn. J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really

measure. Psychological Bulletin, 101, 171-191.

Flynn. J. R. (1999). Searching for justice: The discovery of IQ gains over time. American Psychologist, 54, 5-20.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. Psychological Bulletin, 99 (3), 330-337.

*Fox, H. & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. Personnel Psychology, 27, 209-223.

Frisby, C. L. (1995). Introduction to the Bell Curve commentaries. School Psychology Review, 24, 9-11.

*Gael, S. & Grant, D. L. (1972). Employment test validation for minority and nonminority telephone company service representatives. Journal of Applied Psychology, 56 (2), 135-139.

*Gael, S., Grant, D. L., & Ritchie, R. J. (1975a). Employment test validation for minority and nonminority telephone operators. Journal of Applied Psychology, 60 (4), 411-419.

*Gael, S., Grant, D. L., & Ritchie, R. J. (1975b). Employment test validation for minority and nonminority clerks with work sample criteria. Journal of Applied Psychology, 60 (4), 420-426.

*Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B. & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. Personnel Psychology, 51, 357-374.

Gottfredson, L. S. (Ed.). (1986). The g factor in employment. Journal of Vocational Behaviour, 29 (3).

*Grant, D. L. & Bray, D. W. (1970). Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 54 (1), 7-14.

Guion, R. M. (1998). Assessment, Measurement, and Prediction for Personnel Decisions. Mahwah, N. J.: Lawrence Erlbaum.

Hartigan, J. A. & Wigdor, A. K. (1989). Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.

*Harville, D. L. (1996). Ability test equity in predicting job performance work samples. Educational and Psychological Measurement, 56 (2), 344-348.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press, Inc.

Helms, J. E. (1997). The triple quandary of race, culture, and social class in standardized cognitive ability testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary Intellectual Assessment. New York: The Guilford Press.

Hoffman; C. C., Nathan, B. R. & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. Personnel Psychology, 44, 601-618.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Hunter, J. E. & Schmidt, F. L. (1990). Methods of meta-analysis: correcting error and bias in research findings. California: Sage Publications, Inc.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), Perspectives on Bias in Mental Testing. New York: Plenum.

*Jacobs, R. R., Conte, J. M., Day, D. V., Silva, J. M., & Harris, R. (1996). Selecting bus drivers: Multiple predictors, multiple perspectives on validity, and multiple estimates of utility. Human Performance, 9 (3), 199-217.

Jensen, A. R. (1980). Bias in Mental Testing. New York: The Free Press.

Kraiger, K. & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.

Landy, F. J., Shankster, L. J. & Kohler, S. S. (1994). Personnel selection and placement. Annual Review of Psychology, 45, 251-296.

*Lefkowitz, J. (1972). Differential validity: Ethnic group as a moderator in predicting tenure. Personnel Psychology, 25, 223-240.

*Lopez, F. (1966). Current problems in test performance of job applicants: I. Personnel Psychology, 19, 10-18.

Martocchio, J. J. & Whitener, E. M. (1992). Fairness in personnel selection: A meta-analysis and policy implications. Human Relations, 45 (5), 489-506.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. Personnel Psychology, 41, 283-314.

More normal nonsense. (1989, July). Fortune, p.118.

Nathan, B. R. & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. Personnel Psychology, 41, 517-535.

Neisser, U. (1998). Introduction: Rising test scores and what they mean. In The Rising Curve: Long-Term Gains in IQ and Related Measures. Ulric Neisser (Ed.). American Psychological Association: Washington, DC.

Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51, 77-101.

*Pulakos, E. D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. Human Performance, 9 (3), 241-258.

*Pulakos, E. D., Schmitt, N., & Chan, N. (1996). Models of job performance ratings: An examination of ratee race, ratee gender, and rater level effects. Human Performance, 9 (2), 103-119.

Pulakos, E. D., White, L. A., Oppler, S. H. & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. Journal of Applied Psychology, 74 (5), 770-780.

Ree, M. J. & Earles, J. A. (1991). Predicting training success: Not much more than g. Personnel Psychology, 44, 321-332.

*Roberts, H. E. & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. Military

Psychology, 8 (2), 95-113.

Sackett, P. R. & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. Personnel Psychology, 50, 707-721.

Sackett, P. R. & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. American Psychologist, 49 (11), 929-954.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 58, 5-9.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 33, 705-724

Schmitt, N., Clause, C. S., & Pulakos (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), International Review of Industrial and Organizational Psychology (Vol. 11, pp. 115-139). Chichester, England UK: John Wiley & Sons Ltd.

*Schmitt, N., Hattrup, K., & Landis, R. S. (1993). Item bias indices based on total test score and job performance estimates of ability. Personnel Psychology, 46, 593-611.

Schwarzer, R. (1989). Meta-Analysis Programs (Version 5.0) [Computer software manual]. Berlin, Germany: Author.

Schwarzer, R. (1991). Meta-Analysis Programs (Version 5.3) [Computer software]. Berlin, Germany: Author.

Shepard, L. A. (1987). The case for bias in tests of achievement and scholastic aptitude. In S. Modgil & Celia Modgil (Eds.), Arthur Jensen: Consensus and Controversy (pp. 177-190). Philadelphia, PA.: The Falmer Press, Taylor and Francis Inc.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the validation and use of personnel selection procedures. College Park, MD: Author.

Thorndike, R. L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8 (2), 63-70.

Tucker, W. H. (1994). The Science and Politics of Racial Research. Urbana: University of Illinois Press.

U.S. Department of Labor (1983). Test validation for 12,000 Jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (USES Test Research Rep. No. 45). Division of Counseling and Test Development, Employment and Training Administration. Washington, DC: Author.

*Villanova, P., Bernardin, H. J., Johnson, D. L. & Dahmus, S. A. (1994). The validity of a measure of job compatibility in the prediction of job performance and turnover of motion picture theater personnel. Personnel Psychology, 47, 73-90.

Wigdor, A. K. & Garner, W. R. (1982). Ability testing: Uses, consequences, and controversies. Washington, DC: National Academy Press.

Winston, A. S. (1996). The context of correctness: A comment on Rushton. Journal of Social Distress and the Homeless, 5 (2), 231-250.

Winston, A. S. (1998). Science in the service of the far right: Henry E. Garrett, the IAAEE, and the Liberty Lobby. Journal of Social Issues, 54 (1), 179-210.

*Wright, P. M., Kacmar, K. M., McMahan, G. C. & Deleeuw, K. (1995). P=f(M X A): Cognitive ability as a moderator of the relationship between personality and job performance. Journal of Management, 21(6), 1129-1139.

Appendix

Summary of Meta-Analysis Study Characteristics

Reference	n		Effect size (d)	Job	Measure source	Classification
	White	Black				
Baehr, et al. (1971)	170	62	0.29	Police Patrolman	Generic	Cognitive ability
	170	62	0.08	Police Patrolman	Supervisor rating	Subjective
	170	61	0.29	Police Patrolman	Supervisor rating	Subjective
	167	60	-0.01	Police Patrolman	Tenure	Objective
	170	62	0.33	Police Patrolman	Total awards	Objective
	170	62	0.49	Police Patrolman	Complaints	Objective
	170	62	1.06	Police Patrolman	Disciplinary actions	Objective
	170	62	-1.01	Police Patrolman	# of arrests	Objective
	170	62	-0.09	Police Patrolman	Absence	Objective
Dalldorf & Holmgren (1993)[a]	132	53	0.82	Bkkeep/account clerk	AFQT	Cognitive ability
	132	53	-0.02	Bkkeep/account clerk	Supervisor rating	Subjective
	220	30	0.76	Cosmetologist	AFQT	Cognitive ability
	220	30	0.37	Cosmetologist	Supervisor rating	Subjective
	133	16	1.18	Engine mechanic	AFQT	Cognitive ability
	132	16	0.47	Engine mechanic	Supervisor rating	Subjective
	108	12	0.92	Electronic technician	AFQT	Cognitive ability
	108	12	0.84	Electronic technician	Supervisor rating	Subjective
	237	40	0.54	Firefighter	AFQT	Cognitive ability
	236	40	0.62	Firefighter	Supervisor rating	Subjective
	264	19	0.63	Operating engineer	AFQT	Cognitive ability
	264	19	0.81	Operating engineer	Supervisor rating	Subjective

	60	40	1.29	Computer operator	AFQT	Cognitive ability
	60	40	-0.05	Computer operator	Supervisor rating	Subjective
	152	106	1.12	Lic. practical nurse	AFQT	Cognitive ability
	152	106	0.54	Lic. practical nurse	Supervisor rating	Subjective
	66	88	1.27	Word processing	AFQT	Cognitive ability
	66	88	0.45	Word processing	Supervisor rating	Subjective
DuBois, et al. (1993)[a]	354	105	0.70	Supermarket cashier	Generic	Cognitive ability
	554	89	0.94	Supermarket cashier	Generic	Cognitive ability
	315	82	0.48	Supermarket cashier	Supervisor rating	Subjective
	540	84	0.33	Supermarket cashier	Supervisor rating	Subjective
	155	17	0.21	Supermarket cashier	Speed	Objective
	246	53	0.30	Supermarket cashier	Accuracy	Objective
	421	62	0.53	Supermarket cashier	Speed	Objective
	480	76	0.39	Supermarket cashier	Accuracy	Objective
Farr, et al. (1971)	101	42	0.59	Toll collectors	Arithmetic	Cognitive ability
	111	42	-0.03	Toll collectors	Absence	Objective
	114	43	-0.17	Toll collectors	Termination	Objective
	94	35	0.24	Toll collectors	Dollar accuracy	Objective
	94	35	0.08	Toll collectors	Axle accuracy	Objective
	207	41	0.40	Correctional officer	CTMM	Cognitive ability
	322	49	0.32	Correctional officer	Supervisor rating	Subjective
	322	49	0.29	Correctional officer	Absence	Objective
	319	49	0.31	Correctional officer	Promotion	Objective
	56	18	0.63	Toll facility officer	Otis	Cognitive ability
	56	18	0.15	Toll facility officer	Supervisor rating	Subjective
	55	17	-0.19	Toll facility officer	Promotion	Objective
	51	16	0.08	Toll facility officer	Absence	Objective

	363	46	0.33	Home officer clerical	TMA	Cognitive ability
	306	37	0.43	Home officer clerical	Quickness	Subjective
	296	37	0.45	Home officer clerical	Accuracy	Subjective
	274	32	0.38	Home officer clerical	Numerical ability	Subjective
	306	37	0.41	Home officer clerical	Verbal ability	Subjective
	296	36	0.31	Home officer clerical	Judgement	Subjective
	306	37	0.49	Home officer clerical	Mental ability	Subjective
	296	35	0.52	Home officer clerical	Promotion potential	Subjective
	104	24	0.24	Keypunch operator	TMA	Cognitive ability
	103	28	0.46	Keypunch operator	Concentration	Subjective
	103	28	0.10	Keypunch operator	Learning ability	Subjective
	103	28	0.26	Keypunch operator	Work sharing	Subjective
	103	28	0.30	Keypunch operator	Error detection	Subjective
	103	28	0.41	Keypunch operator	Social interaction	Subjective
	103	28	0.19	Keypunch operator	Overall effective.	Subjective
	79	21	-0.10	Keypunch operator	Keypunch speed	Objective
	76	18	-0.32	Keypunch operator	Error %	Objective
Fox & Lefkowitz (1974)	67	77	0.18	Factory manufacturer	SRA non-verbal	Cognitive ability
	67	77	-0.05	Factory manufacturer	Supervisor rat./rank	Subjective
	67	77	-0.28	Factory manufacturer	Efficiency	Objective
Gael et al. (1975a)	464	501	0.76	Telephone operator	BSQT	Cognitive ability
	464	501	0.62	Telephone operator	Work samples	Subjective
Gael et al. (1975b)	185	143	0.76	Clerk	BSQT	Cognitive ability
	185	143	0.84	Clerk	Work samples	Subjective
Gael & Grant (1972)	193	106	0.71	Tel. service rep.	BSQT	Cognitive ability
	193	107	0.19	Tel. service rep.	Verbal contract	Subjective

	193	107	0.25	Tel. service rep.	Filing	Objective
	193	107	0.20	Tel. service rep.	Record prep.	Objective
Goldstein, et al. (1998)[a]	545	88	0.83	Managerial	Wesman Personnel Classification Test	Cognitive ability
	545	88	0.41	Managerial	Assessment center	Subjective
	545	88	-0.12	Managerial	Supervisor rating	Subjective
Grant & Bray (1970)	219	211	0.30	Tel. install & repair	SCAT	Cognitive ability
	219	211	0.28	Tel. install & repair	Training lvl. passed	Objective
Harville (1996)[a]	81	35	0.53	Information systems radio operator	AFQT	Cognitive ability
	81	35	-0.06	Information systems radio operator	Work sample	Subjective
	106	53	0.36	Personnel	AFQT	Cognitive ability
	106	53	0.02	Personnel	Work sample	Subjective
	112	40	0.29	Aircrew life support	AFQT	Cognitive ability
	112	40	-0.04	Aircrew life support	Work sample	Subjective
Jacobs, et al. (1996)[a]	299	417	0.47	Bus driver	Generic	Cognitive ability
	220	365	0.05	Bus driver	Dependability	Subjective
	220	365	0.07	Bus driver	Schedule adherence	Subjective
	220	365	0.01	Bus driver	Accidents/Safety	Subjective
	220	365	0.14	Bus driver	Drive quality	Subjective
	220	365	-0.06	Bus driver	Attention to details	Subjective
	220	365	-0.12	Bus driver	Rider interactions	Subjective
	220	365	0.02	Bus driver	Service orientation	Subjective
	220	365	-0.12	Bus driver	Supervisor interaction	Subjective

	220	365	-0.05	Bus driver	Peer interaction	Subjective
	293	414	-0.02	Bus driver	Absence	Objective
	294	414	0.00	Bus driver	Accidents	Objective
Lefkowitz (1972)	164	190	0.18	Factory manufacturer	SRA non-verbal	Cognitive ability
	255	289	-0.18	Factory manufacturer	Job tenure	Objective
Lopez (1966)	80	102	0.62	Toll collector	Generic	Cognitive ability
	36	56	-0.02	Toll collector	Supervisor rating	Subjective
	80	102	0.23	Toll collector	Absence	Objective
	80	102	0.29	Toll collector	Accuracy	Objective
Pulakos & Schmitt (1996)	259	100	1.25	Federal investigative	AFQT	Cognitive ability
Pulakos, et al. (1996)	259	100	0.44	Federal investigative	Work sample	Subjective
(Articles use same data)	259	100	0.99	Federal investigative	Work sample	Subjective
	259	100	0.26	Federal investigative	Supervisor rating	Subjective
	259	100	0.85	Federal investigative	Written	Subjective
	259	100	0.58	Federal investigative	Role Play	Subjective
Roberts & Skinner (1996)	12 453	511	0.66	Military officer	AFOQT	Cognitive ability
	12 453	511	0.35	Military officer	Course grade	Subjective
	12 453	511	0.08	Military officer	Effectiveness rpt.	Subjective
Schmitt et al. (1993)	207	32	0.99	Various trades	Generic	Cognitive ability
	207	32	0.68	Various trades	Work sample	Subjective
Villanova, et al. (1994)[a]	154	37	1.24	Motion picture theater staff	Generic	Cognitive ability
	154	37	1.34	Motion picture theater staff	Supervisor rating	Subjective

	154	37	0.19	Motion picture theater staff	Turnover	Objective
Wright, et al. (1995)[a]	262	49	0.91	Warehousers	Wonderlic+Generic	Cognitive ability
	206	38	0.07	Warehousers	Supervisor rating	Subjective

Note. To save space, "et al." is used for studies with three or more authors. The GATB validation database is not included given its extensive size. n = sample sizes used to calculate effect size. d = unadjusted effect size (Hedges & Olkin, 1985). AFQT = Armed Forces Qualification Test; Generic = Custom test for study/job; CTMM = California Test of Mental Maturity; TMA = Thurstone Test of Mental Alertness; SRA non-verbal = Science Research Associates' nonverbal intelligence test; BSQT = Bell System Qualification Test; SCAT = School and College Ability Test; AFOQT = Air Force Officer Qualifying Test.

[a]Researcher(s) contacted to provide additional data.